Title

Bond-Order Time Series Analysis for Detecting Reaction Events in Ab Initio Molecular Dynamics Simulations.

Journal

Journal of chemical theory and computation, 16(3)

ISSN

1549-9618

Authors

Hutchings, Marshall
Liu, Johnson
Qiu, Yudong
et al.

Publication Date

2020-03-01

DOI

10.1021/acs.jctc.9b01039

Bond order time series analysis for detecting reaction events in ab initio molecular dynamics simulations

Marshall Hutchings,[†] Johnson Liu,[†] Yudong Qiu,[†] Chenchen Song,[‡] and Lee-Ping Wang[*,†]

†*Department of Chemistry, University of California; 1 Shields Ave; Davis, CA 95616.*

‡*Department of Chemistry, Stanford University; Stanford, CA 94305.*

¶*SLAC National Accelerator Laboratory; Menlo Park, CA 94025.*

E-mail: leeping@ucdavis.edu

Abstract

Ab initio molecular dynamics is able to predict novel reaction mechanisms by directly observing the individual reaction events that occur in simulation trajectories. In this article, we describe an approach for detecting reaction events from simulation trajectories using a physically motivated model based on time series analysis of *ab initio* bond orders. We found that applying a threshold to the bond order was insufficient for accurate detection, whereas peak finding on the first time derivative resulted in significantly improved accuracy. The model is trained on a reference set of reaction events representing the ideal result given unlimited computing resources. Our study includes two model systems: a heptanylium carbocation that undergoes hydride shifts, and an unsaturated iron carbonyl cluster that features CO ligand migration and bridging behavior. The results indicate a high level of promise for this analysis approach to be used in mechanistic analysis of reactive AIMD simulations more generally.

1 Introduction

A central goal of theoretical chemistry is to provide sufficient insight into reactivity at the molecular scale to inform the design of experiments including reaction routes, reaction conditions, and catalysis.[1–3] Computational studies of reaction mechanisms often start by hypothesizing a reaction pathway from chemical intuition, followed by calculating the minimum energy path and associated critical points (reactant, product, and transition state structures) with local optimization methods.[4,5] The reaction rate associated with a pathway may be estimated from the activation energy using kinetic models, enabling a semi-quantitative comparison with experiment.[6,7] The main drawback of this strategy is that only existing hypotheses can be tested, and such hypotheses traditionally originate from chemical intuition; in other words, the systematic generation of mechanistic hypotheses is an important challenge for theoretical chemistry. Another aspect of this challenge is that many reaction mechanisms proceed through multiple elementary steps and short-lived intermediates that are difficult to experimentally characterize.

Recently, computational methods have been developed that automate the searching procedure by systematically applying basic rules to break and form chemical bonds in a combinatorial fashion.[8–15] These methods, which are based on assuming general rules of reactivity rather than specific mechanistic hypotheses, can greatly increase the automation in mechanistic studies and have proven successful in applications.[16–18] However, there are still limitations to such approaches because they require assuming the basic rules of reactivity, which are not fully understood; moreover, the relative positioning of reactants in multi-molecular or roaming reaction pathways continues to be a challenge for rules-based approaches.

In the past few years, *ab initio* molecular dynamics (AIMD) has emerged as a useful tool for the discovery of reaction mechanisms. In fact, classical molecular mechanics (MM) simulations have long been used to discover pathways of protein folding and conformational change;[19–23] these involve changes in the protein backbone and side chain conformations as well as intermolecular interactions, which do not require a quantum mechanical description.

Consequently, the PES can be approximated using inexpensive force fields, allowing MM simulations to routinely reach microsecond time scales and beyond. On the other hand, predictive sampling of a reactive system usually requires a quantum mechanical calculation of the electronic wavefunction at every time step, which costs at least four orders of magnitude more than evaluating a MM force field and usually scales less favorably with system size. More recently, modern advances in electronic structure methods and accelerated hardware implementations have resulted in speed-ups of 2-3 orders of magnitude for Hartree-Fock and density functional theory (DFT) calculations,[24–35] placing AIMD simulations on the threshold of discovering reaction mechanisms that occur on nanosecond or longer timescales.[36–41]

Because reaction rates are exponentially decreasing functions of the activation barrier, it is still highly challenging to map the chemically interesting reaction pathways in an unbiased AIMD simulation. Recently, we and others have introduced specialized AIMD simulation methods for accelerating the discovery of reaction pathways. The Pietrucci group introduced topological-based permutation invariant "SPRINT" coordinates helped to address the isomer degeneracy problem in metadynamics.[42] The Pfaendtner group demonstrated how to reduce computational cost and the need to manually specify reaction coordinates by using parallel bias metadynamics using SPRINT coordinates as collective variables.[43] The *ab initio* nanoreactor causes a large number of reactions to occur in a relatively short simulation by periodically forcing the molecules in the simulation to undergo high-velocity collisions.[36] Because the nanoreactor requires no specification of reaction coordinate, it is able to discover new pathways for interesting reactions such as the prebiotic synthesis of glycine and sugars.[36,44,45] As these simulations do not involve specifying reaction coordinates or desired products, an automatic approach is needed to identify the potentially interesting reaction events.

The recent emergence of AIMD simulations containing large numbers of reaction events requires new theoretical tools for deriving useful knowledge from them. One of the principal tasks is to identify the discrete transition between chemical structures from the continuous

variables of the simulation trajectory. We recently introduced a procedure for detecting and extracting reaction events based on analysis of interatomic distances, followed by a series of optimization calculations to locate the minimum energy path associated with the observed reaction.[46] In a related work, Döntgen and coworkers developed an analysis approach for reactive MD trajectories simulated using the ReaxFF force field, where the ReaxFF bond order was used to detect reaction events and calculate reaction rates directly from the observed events.[47] Both studies noted that some ambiguity remains in reaction event detection, as a number of empirical parameters (including covalent radii, ReaxFF parameters, and lag times) were used to determine the threshold for what constituted a genuine reaction event in the simulation. As these exploratory-type simulations are destined to become increasingly important in simulation studies of reaction mechanisms, a greater amount of rigor and precision is clearly needed in the identification of the reaction events. Motivated by this need, we would like to address the following questions: How can we properly define a reaction event in an AIMD simulation? How can we systematically improve on reaction event detection methods?

In this paper, we address these questions by introducing a new reaction event detection method based on time series analysis of the AIMD trajectory. To develop this method, a suitable set of reference reaction events is created by local energy minimization of each structure on a reactive trajectory. The sequence of optimized structures is clustered into a discrete number of chemical states, and the transitions in the sequence of states are used as a reference dataset for the time series analysis. Our reaction detection approach is based on the *ab initio* bond order index defined by Mayer.[48] Our results show that the time series analysis based on bond order indices is able to reproduce the reference data set accurately using few parameters. An iron carbonyl cluster ($Fe_3(CO)_9$) previously studied theoretically by Schaefer and coworkers[49] and a heptanylium cation ($C_7H_{15}{}^+$) are used as a testing ground for this method; our results indicate the AIMD simulation method is able to discover a significant number of new local minima connected by low energy barriers at a far lower cost

than optimizing entire trajectories. Our methods and results provide a foundation for more chemically relevant understanding of reactive AIMD trajectories.

2 Theory and Methods

Here we briefly summarize the main considerations in the development of our reaction detection model before describing individual aspects in more detail. This paper focuses on reactions that involve rearrangements of bonding within a single molecule, though we think making generalizations to reactivity involving several molecules should be conceptually straightforward. Because our reaction events involve making and/or breaking chemical bonds, we intuitively expect the atom pair-wise bond orders (BO) will increase or decrease when bonds are formed, broken, or undergo changes in electronic character. Thus, our model will use the BO time series between all atom pairs as input data and detect reaction events from changes in the time series. We use the *ab initio* bond order defined by Mayer as:

$$M_{ab}[i] = 2 \sum_{\mu \in a} \sum_{\nu \in b} \left[(\mathbf{P}^{\alpha}\mathbf{S})_{\mu\nu} (\mathbf{P}^{\alpha}\mathbf{S})_{\nu\mu} + (\mathbf{P}^{\beta}\mathbf{S})_{\mu\nu} (\mathbf{P}^{\beta}\mathbf{S})_{\nu\mu} \right] [i] \tag{1}$$

where $\mathbf{P}^{\alpha,\beta}$ is the one-particle density matrices for alpha and beta spin, $\mathbf{S}$ is the overlap matrix, μ, ν are indices for atomic basis functions, the sums are restricted to functions centered on atom indices a, b, and $[i]$ indicates values at frame i in the simulation trajectory. Thus, the bond order is defined as a discrete series spaced in time by the simulation time step δ.

In the context of our work, "detection" refers to estimating or predicting the approximate location of a reaction event. This definition requires introducing a set of reference reaction events that represents the desired result given unlimited computing resources. A reference reaction event is defined when the *ab initio* molecular dynamics trajectory crosses between two catchments (energy basins) in configuration space that contain chemically different local minima. Energy basins are separated by manifolds of local maxima (dividing surfaces), and

we assume the minima are located in the interior of the catchment and away from dividing surfaces, such that chemically different energy-minimized species will differ significantly in their structures and BO matrices. Therefore, if we could carry out energy minimization of every trajectory frame, the chemically distinct species and reaction events could be precisely located by comparing the BO matrices of energy-minimized structures (Figure 1).

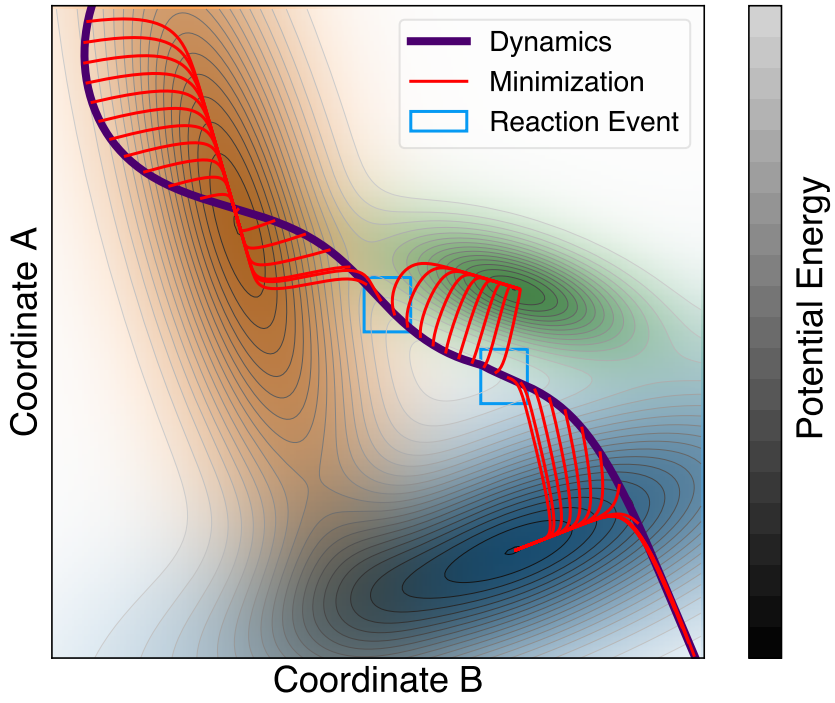

Figure 1: Example MD trajectory (violet) with the optimization pathways for the discrete time steps shown in red. Light blue boxes highlight where the trajectory crosses subdomains in configuration space. These crossings are defined as reaction events.

The reference method is too computationally costly for routine applications because energy minimization of every trajectory frame is significantly more expensive than the AIMD simulation itself. Here, we have computed the reference reaction events to train the model parameters for our two systems, the iron carbonyl cluster $Fe_3(CO)_9$ and heptanylium cation $C_7H_{15}{}^+$; these systems have major differences in terms of their composition, bonding and

coordination. By applying our method to both systems and comparing the results, we characterize the parameter sensitivity of the reaction detection model and provide some guidelines for when it is necessary to compute the reference reaction events for a system of interest.

2.1 Computational Details

To generate a set of reference reaction events and bond order time series for both systems we used unbiased, temperature-accelerated *ab initio* molecular dynamics simulations.[50,51] For both systems we used a Velocity Verlet integrator with a timestep of 1 fs and a Langevin thermostat with an equilibrium temperature of 1000 K and a damping time of 1 ps^{-1}. We simulated the iron carbonyl cluster using the BP86 density functional approximation together with a double-ζ plus polarization (DZP) all-electron basis for all atoms including iron, following Ref.[49] The molecular dynamics simulation was propagated for a duration of 8,373 steps before terminating with a SCF convergence error. The heptanylium simulation used the B3LYP density functional and a 6-31G* basis set, and the simulation was propagated for 10,000 steps. To create the reference reaction sets, every AIMD frame was used as the input coordinates for energy minimization at the same level of theory.[52] All of the simulations in this study were carried out using the TeraChem quantum chemistry software package.[24–26]

2.2 Details of the model systems

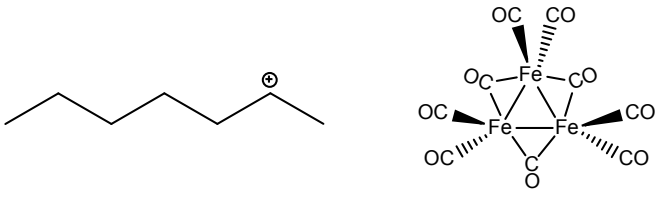

Figure 2: Starting structures of the two systems; heptanylium cation $C_7H_{15}{}^+$ (left) and iron carbonyl cluster $Fe_3(CO)_9$(right).

We chose two model systems to characterize the accuracy of our reaction detection model in this study. The first system with simpler and more straightforward reactivity is the heptanylium alkyl carbocation, $C_7H_{15}{}^+$. We expected the reaction events for the heptanylium system to manifest as hydride and methyl shifts. This system was chosen so that generating a tertiary carbocation was unlikely under our simulation conditions, as this would halt further reactions because of their relative stability.

The iron carbonyl cluster we chose to study is $Fe_3(CO)_9$ which is the smallest in a series of four clusters with increasing carbonyl count studied by Schaefer and coworkers.[49] Because $Fe_3(CO)_9$ is unsaturated, this system presents interesting possibilities for CO ligand migration and bridging multiple Fe atoms. These two systems were chosen to be chemically distinct in order to illustrate the performance of the model when used in diverse applications.

2.3 Reference reaction events

We assume that the potential energy surface is divided into catchments or energy basins denoted as $\mathcal{S}_k$ in the regions of the potential energy surface accessed by the AIMD simulation, where the index k represents all such basins that are sampled by one simulation. These are bounded regions on the potential energy surface where each point in the region is mapped by energy minimization to a local minimum somewhere in the interior as $\mathbf{y}_k = \mathit{Optimize}\,(\mathbf{x} \in \mathcal{S}_k)$. Moreover, because we are interested in detecting reactivity, catchments that correspond to chemically identical species and share any boundaries are grouped together. Our task consists of finding the catchments that are visited by the AIMD trajectory frames and identifying when the trajectory crosses over their dividing surfaces (i.e. reaction events).

We expect that two local minima in different energy basins $(\mathbf{y}_k, \mathbf{y}_l)$ with major differences in chemical bonding should be distinguishable by comparing their BO matrices. Thus, constructing the reference reaction events from an AIMD trajectory follows this procedure:

1. Calculate a series of optimized structures by local energy minimization of every frame

in the simulation trajectory, i.e. $\mathbf{y}[i] = Optimize(\mathbf{x}[i])$.

2. Cluster the series of optimized structures using a chosen distance metric and clustering algorithm. This produces a set of clusters $\{C_k, 1 \leq k \leq N_C\}$ where each trajectory frame belongs to only one cluster and N_C is the number of clusters. Each cluster k corresponds to a distinct catchment $\mathcal{S}_k$ and a representative optimized structure $\mathbf{y}_k$. The cardinality of the cluster is represented as $|C_k|$.

3. Assign each optimized structure to a cluster to produce a series of cluster numbers $\{K[i], 1 \leq i \leq N_{\text{steps}}\}$.

4. The time coordinates of reference reaction events are where the cluster number of the optimized structure differs between two consecutive frames as:

$$\mathcal{E}_{\text{ref}} = \{i \mid K[i] \neq K[i+1]\} \tag{2}$$

For two energy-minimized structures, we compute the bond-order distance metric (BODM) as the L_2 norm of the difference in bond order matrices:

$$d[i,j] = \sqrt{\sum_{a<b}^{N_{\text{atom}}} \left(\widetilde{M}_{ab}[i] - \widetilde{M}_{ab}[j]\right)^2} \tag{3}$$

where the tilde over $\widetilde{M}$ indicates that the BO matrix of the energy-minimized structure is used. This idea is similar to, and indeed inspired by, the featurization of biomolecular simulation trajectories such as contact maps, dihedral angles, and metrics such as RMSD which are used in the construction of kinetic models.[53,54] Our choice of using BO matrices is an important distinguishing factor from earlier work, and justified because the BODM directly measures changes in chemical bonding and should exclude other conformational changes. (*Remark:* Two structures that differ only by the permutation of atomic indices may also have significant BODMs. This does not significantly affect our main conclusions.)

To cluster the optimized structures into discrete chemical species, we used a hierarchical clustering algorithm with an average-linkage criterion as implemented in the SciPy package,[55,56] where the input data is a matrix of the BODM values above. At the start of the algorithm, each structure is assigned to a separate cluster. The pair of clusters with the smallest distance is merged into a single cluster, and the new clusters are renumbered in consecutive ascending order. The distance between the newly merged cluster C_k and all other clusters C_l is defined as:

$$D_{kl} \equiv \frac{1}{|C_k||C_l|} \sum_{i \in C_k,\, j \in C_l} d[i,j] \tag{4}$$

The merging procedure is repeated until the smallest pairwise distance is larger than a threshold parameter, resulting in the final set of clusters $\{C_k\}$. We chose a clustering threshold parameter of 1.0 because it represents a difference of approximately one bond between clusters, and because the number and contents of clusters was consistent with our chemical intuition and visual examination of the optimized structures. A dendrogram showing the successive merging of clusters as a function of the threshold parameter for each system is given in Supporting Figures S1 and S2.

A sequence of cluster indices is obtained for the trajectory of optimized frames. For each frame where the cluster number differs between the current and next frame, a reference reaction event is defined. Each reference reaction event is a data structure containing the current frame number, as well as the optimized structures and BO matrices of the current and next frame. The BO matrices allow us to query which bonds were formed or broken in the reaction event, which will become important in §2.5.

2.4 Reaction detection by time series analysis

Here we describe efficient and approximate models for estimating the reaction events via direct analysis of the AIMD BO trajectory data. The purpose of these models is to reduce

the number of computationally costly energy minimizations needed to find the reaction events in the simulation. In our current context where the entire system consists of a single molecule, the model predicts which time coordinates (i.e. frame numbers) are likely to be near true reaction events, thereby restricting the energy minimizations to within small time windows of these predicted frames. A high-quality model should be sensitive enough to correctly detect most or all of the reaction events, while ruling out "unreactive" parts of the simulation trajectory to reduce computational cost.

For a particular atom pair with indices a and b, the bond order time series $\{M_{ab}\} = \{M_{ab}[i]; 1 \leq i \leq N_{\text{steps}}\}$ is a discrete sampling of the bond order as a function of time. Because variations in the bond orders are slow compared to the time step, we assume aliasing effects from discrete sampling are negligible. $\{M_{ab}\}$ contains both long-lasting changes that represent genuine reaction events and changes in chemical bonding, as well as higher-frequency fluctuations that we are less interested in. Thus, we process $\{M_{ab}\}$ with a low-pass filter to remove the fast fluctuations and retain the chemically important features of the time series:

$$\{\overline{M}_{ab}(\sigma)\} = L\Big(\{M_{ab}\},\ \sigma\Big) \tag{5}$$

Here, L is the function that performs the low-pass filtering (we used a sixth-order Butterworth filter), the line over $\overline{M}$ indicates that the time series has been smoothed, and σ represents the cutoff frequency parameter. σ can be optimized in order to produce the best agreement between the detected reaction events and the reference set. One of our goals in this paper is to show that the performance of this method is not highly sensitive to the choice of σ for different applications.

2.4.1 Thresholding on time series values

One intuitive approach to predicting reaction events is to detect when the smoothed time series crosses over a threshold that separates bonded from non-bonded regimes. This approach is similar to our previous study[46] where connectivity between atom pairs was defined by comparing interatomic distances to a threshold derived from covalent radii. One advantage of using bond orders is that the highly sensitive element-wise radius parameters are no longer needed. Here we will show that applying a threshold to the bond order is insufficient for detecting reactions, which motivates the time derivative approach in § 2.4.2.

Equation 6 is the set of predicted reaction events for atom pair (a, b) where $\overline{M}_{ab}$ crosses a threshold μ:

$$\mathcal{E}_{0;ab}(\sigma, \mu) = \left\{ i \;\middle|\; \overline{M}_{ab}(\sigma)[i] > \mu > \overline{M}_{ab}(\sigma)[i+1] \vee \overline{M}_{ab}(\sigma)[i] < \mu < \overline{M}_{ab}(\sigma)[i+1] \right\} \tag{6}$$

The set of predicted reaction events for the entire system is found by taking the union over all atom pairs:

$$\mathcal{E}_0 \;=\; \bigcup_{b>a=1}^{N_{\text{atoms}}} \mathcal{E}_{0;ab} \tag{7}$$

However, we found that the reaction events identified in Equation 7 were incomplete, as many reaction events could not be accurately predicted using a single threshold in the iron carbonyl simulation.

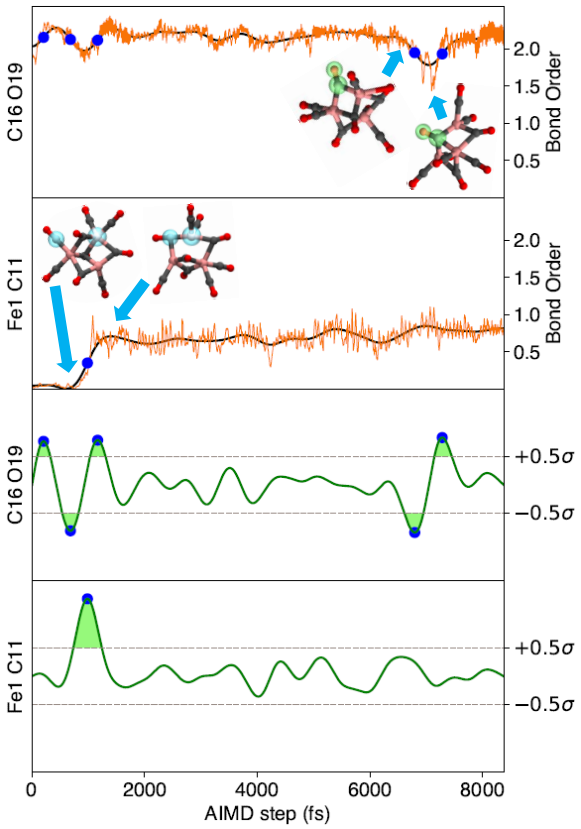

Figure 3: Top two panels: Raw (M_{ab}, orange) and 40 cm^{-1} low-pass filtered ($\overline{M}_{ab}$, black) time series of two selected bond orders. Bottom two panels: First time derivative of filtered series ($\overline{M}'_{ab}$, green curve) with threshold $\mu = \pm 0.5\sigma$ (dashed lines). $\text{Intvl}_{+;ab}$ and $\text{Intvl}_{-;ab}$ shown with light green shading. Detected reaction events $\mathcal{E}_{1;ab}$ from the bottom two panels are shown as blue dots across all panels. Molecule color scheme: iron: pink, carbon: gray, and oxygen: red.

Figure 3 shows why applying a single threshold to the smoothed bond order time series to detect reaction events can be challenging. In the upper panel showing the Fe-C bond order there is a distinct increase from 0.0 to 0.9 near $t = 1000$ fs, indicating that a threshold parameter of $0.1 - 0.8$ would work well for this atom pair. However, the upper C-O panel contains fluctuations in the bond order near $t = 7000$ fs that are indicative of changes in the carbonyl ligand coordination to Fe, where the value of the bond order is consistently in the $1.9 - 2.3$ range. In order to detect any reaction events in the C-O time series, the threshold would need to be much higher, around 2.0. If we were forced to use different thresholds

for different kinds of bonds, then it would detract from the usefulness of the *ab initio* bond order as a simple criterion for detecting reactivity.

Another drawback of applying a threshold directly to $\{\overline{M}_{ab}(\sigma)\}$ is the risk of detecting large numbers of false positives and false negatives. If the thresholds were chosen close to the mean value of an oscillating BO time series, repeated crossings over the threshold could cause many false positives. Although the number of oscillations may be reduced by increasing the smoothing, it does not address the fundamental problem that the threshold parameter is close to the mean value of the oscillation. In both upper panels of Figure 3, there exist ranges of the threshold parameter that would contain many crossings due to oscillations around an apparent mean. The risk of excessive false positives due to repeated threshold crossings and false negatives due to missed crossings indicates that if we applied a threshold to the bond order to detect reaction events, the results would be highly sensitive to parameter choice, which negatively affects the utility of the method. In what follows, we show that applying a similar thresholding approach to the bond order *time derivative* is a simple way to address many of these issues.

2.4.2 Peak finding on first time derivative

Once the raw time series has been filtered to remove high-frequency components (Equation 5), the first time derivative of the smoothed time series is taken:

$$\overline{M}'_{ab}(\sigma)[i] \equiv \frac{d}{dt}\Big(\overline{M}_{ab}(\sigma)[i]\Big) \approx \frac{M_{ab}(\sigma)[i+1] - M_{ab}(\sigma)[i-1]}{2\delta} \tag{8}$$

The prime on $\overline{M}'_{ab}$ indicates the first time derivative, approximated via central difference on the discrete values of $\overline{M}_{ab}$. Next, a threshold (μ) is applied to $\overline{M}'_{ab}(\sigma)$:

$$\text{Intvl}_{+;ab} \equiv \left\{ (u,v) \;\middle|\; \overline{M}'_{ab}(\sigma)[t] > \mu \;\forall\; t \in (u,v) \wedge \overline{M}'_{ab}(\sigma)[u] \leq \mu \wedge \overline{M}'_{ab}(\sigma)[v] \leq \mu \right\} \quad (9)$$

$$\text{Intvl}_{-;ab} \equiv \left\{ (u,v) \;\middle|\; \overline{M}'_{ab}(\sigma)[t] < -\mu \;\forall\; t \in (u,v) \wedge \overline{M}'_{ab}(\sigma)[u] \geq -\mu \wedge \overline{M}'_{ab}(\sigma)[v] \geq -\mu \right\} \quad (10)$$

Here, $\text{Intvl}_{+;ab}$ and $\text{Intvl}_{-;ab}$ are sets of continuous time intervals for which $\{\overline{M}'_{ab}(\sigma)\}$ is above $+\mu$ and below $-\mu$ respectively. We then collect the time-coordinates of the positive maxima of $\{\overline{M}'_{ab}(\sigma)\}$ above $+\mu$ and the negative minima below $-\mu$ as:

$$\begin{aligned} \mathcal{E}_{1;ab}(\sigma,\mu) \equiv & \left\{ t \;\middle|\; \underset{t\in(u,v)}{\arg\max}\; M'_{ab}(\sigma)[t] \;\forall\; (u,v) \in \text{Intvl}_{+;ab} \right\} \\ & \bigcup \left\{ t \;\middle|\; \underset{t\in(u,v)}{\arg\min}\; M'_{ab}(\sigma)[t] \;\forall\; (u,v) \in \text{Intvl}_{-;ab} \right\} \end{aligned} \quad (11)$$

As a result, we obtain $\mathcal{E}_{1;ab}(\sigma,\mu)$ as the final set of reaction events derived from the BO time derivative for atom pair ab. The smoothing, derivative, and thresholding steps are illustrated in Figure 3. The time-coordinate of every blue dot (identified in the bottom two panels) represents the set of detected reaction events $\mathcal{E}_{1;ab}(\sigma,\mu)$ for the given atom pair ab. Figure 3 shows the advantage of using BO time derivatives instead of applying a threshold directly on the BO values, because the derivative approach can detect reaction events from both the Fe-C and C-O time series whereas the same direct threshold cannot be used for both atom pairs. The fundamental assumption of this approach is that there are no reaction events that change the BO time series very slowly, as that would not be detected by the threshold. We expect this assumption to be generally valid due to the relatively short distance ranges over which chemical bonds are broken and formed, and the atomistic forces along the reaction pathway would prevent bond orders from changing very slowly.

Comparisons to the reference set of reaction events can be made in two ways: either on an "bond-wise" basis, or on an "unified" basis where we take the union over all atom pairs. Thus, the unified set of reaction events is taken by collecting all reaction events for all atom

pairs in the system as:

$$\mathcal{E}_1 \;=\; \bigcup_{b>a=1}^{N_{\text{atoms}}} \mathcal{E}_{1;ab} \tag{12}$$

2.5 Receiver operating characteristic objective function

If our model were perfectly accurate, then for every reaction event predicted, the pair of structures preceding and following the event will minimize to chemically different structures, enabling us to carry out further studies such as reaction pathway optimizations. Because the predictor has imperfect accuracy, the predicted event is generally not identical to the actual event, and the pair of frames corresponding to the current and next trajectory frame will minimize to chemically identical structures. Thus, we should quantify the accuracy of our predictions using some measure of distance to the actual reaction events, or equivalently, by the amount of computational cost it requires to find the actual reaction events starting from the predicted ones. Because we have computed $\mathcal{E}_{\text{ref}}$ in § 2.3, our goal is to optimize the parameters and characterize the accuracy of $\mathcal{E}_1$, thus enabling its application with more confidence in future applications where we do not have $\mathcal{E}_{\text{ref}}$.

In the ideal case, the set of detected reaction events and actual events are equal, and the complete set of reactant and product structures could be found by two energy minimizations for each detected reaction event, with a computational cost of $2 \cdot |\mathcal{E}_{\text{ref}}| \ll N_{\text{steps}}$ times the cost of a single energy minimization. On the other hand, if the predicted reaction event is located close in time to the actual event, then it could be found by energy minimizing more structures in a time window of increasing size around the detected event. As the time window around each element of $\mathcal{E}_1$ is increased (both forwards and backwards in time), an increasing number of true reaction events will be found, and the computational cost is increased as well. In the limiting case, the window size is equal to the entire trajectory length, and all of the reaction events in $\mathcal{E}_{\text{ref}}$ are found at a cost equal to computing $\mathcal{E}_{\text{ref}}$ itself. Thus, the detection method is deemed to be useful if it detects a greater fraction of reaction events in

$\mathcal{E}_{\text{ref}}$ than the fraction of the trajectory that is optimized with a chosen value of the window size. By increasing the window size over a range $(0, N_{\text{steps}})$, we can interpolate between these two limits and construct a receiving operator characteristic (ROC) objective function.

The ROC is a commonly used statistical approach for evaluating the diagnostic ability of a binary classifier, created by plotting the true positive rate (TPR) vs. the false positive rate (FPR) as a sensitivity threshold is varied.[57,58] In our definition of the ROC, we use a time window of variable size $n \cdot \delta$ representing the number of trajectory frames being optimized in the neighborhood of each detected reaction event in $\mathcal{E}_1$. The smallest possible set of optimized frames $\mathcal{X}^{(0)}$ corresponds to a window size of zero:

$$\mathcal{X}^{(0)} = \mathcal{E}_1 \tag{13}$$

where the superscript on $\mathcal{X}$ is the window size.

For any window size w, the set of frames being minimized $\mathcal{X}^{(w)}$ may be defined as:

$$\mathcal{X}^{(w)} \;=\; \left\{ i + n \;\middle|\; i \in \mathcal{X}^{(0)}, -w \leq n \leq w \right\} \cap \mathcal{T} \tag{14}$$

$\mathcal{X}^{(w)}$ is then used to determine the true positive rate ($\text{TPR}^{(w)}$) and false positive rate ($\text{FPR}^{(w)}$). $\text{TPR}^{(w)}$ is the amount of reference reaction events in $\mathcal{E}_{\text{ref}}$ included in the optimized frames $\mathcal{X}^{(w)}$ divided by the total number of reference reaction events $|\mathcal{E}_{\text{ref}}|$. $\text{FPR}^{(w)}$ is calculated as the fraction of trajectory frames not containing reaction events included in $\mathcal{X}^{(w)}$. These functions are defined as:

$$\text{TPR}^{(w)} = \frac{\left|\mathcal{X}^{(w)} \bigcap \mathcal{E}_{\text{ref}}\right|}{|\mathcal{E}_{\text{ref}}|}; \quad \text{FPR}^{(w)} = \frac{\left|\mathcal{X}^{(w)} - \mathcal{E}_{\text{ref}}\right|}{|\mathcal{T} - \mathcal{E}_{\text{ref}}|} \tag{15}$$

The parametric curve $(\text{FPR}^{(w)}, \text{TPR}^{(w)})$ is traced out as w increases, and the ROC objective function $\phi(\sigma, \mu)$ is calculated as the area under the parametric curve as shown in Figure 4.

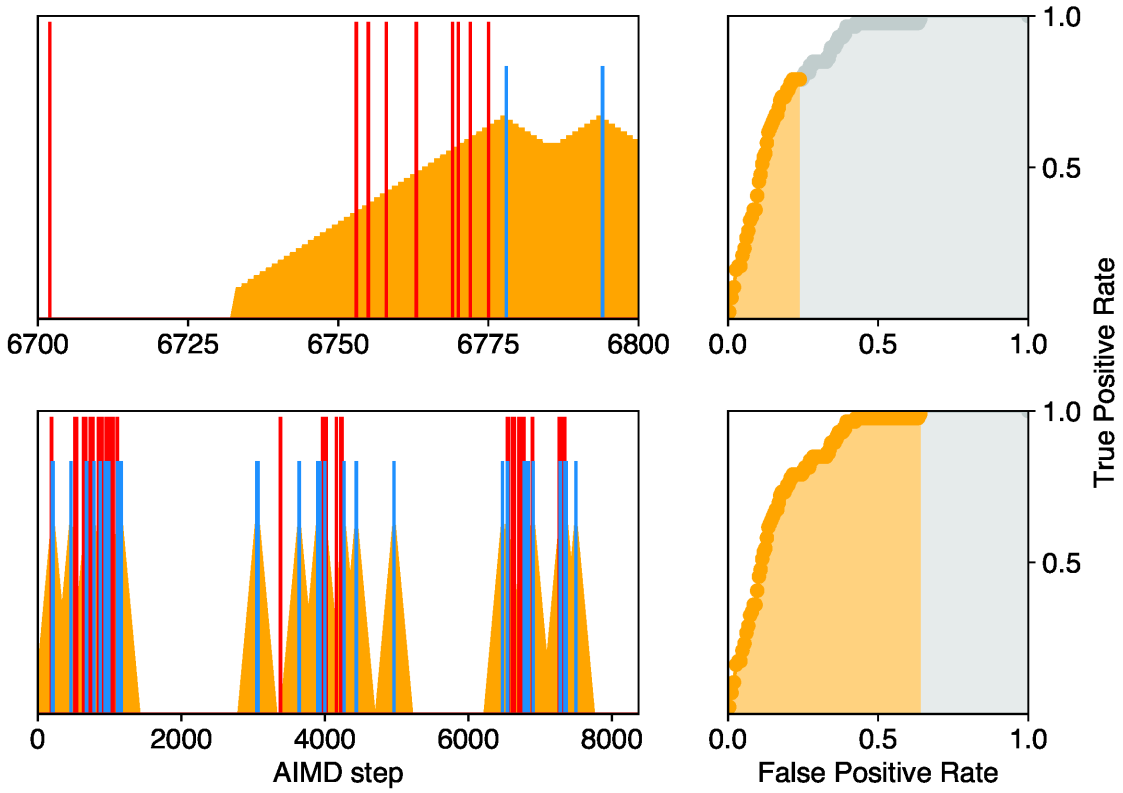

Figure 4: Left panels: Reference reaction event locations (red), BOTS reaction event locations (blue), broadening time windows (orange wedges) around BOTS reaction events. Top panels: Magnification of a 100-frame sequence of the trajectory showing the expanding time window. Bottom panels: Entire trajectory with sufficiently large window size for maximal true positive rate. Right panels: Plot of $\mathrm{TPR}^{(w)}$ vs. $\mathrm{FPR}^{(w)}$ (Equation 15) where orange region indicates the current value of w. The area under the whole curve is the ROC objective function.

The ROC score has an upper bound of 1.0 corresponding to perfect accuracy, i.e. all of the true reaction events are found using a window size of zero, whereas scores of 0.5 or lower indicate the method has no predictive power beyond a random number generator.

2.6 Bond-wise criterion for reaction detection

The procedure defined above uses a unified set of detected reaction events across all atom pairs (Equation 12) to predict the total set of reaction events in the entire system. This

approach was found to be problematic because it ignores the local character of reaction events, i.e. a single reaction event involves changes in bond order for a particular subset of atom pairs. If the predicted reaction event could be mapped to an actual event where different bonds are broken or formed, then it would not be possible to identify the reactive sites within the system; this would become an important deficiency of the method for systems that contain multiple molecules. To resolve this issue, we defined a "bond-wise" criterion that ensures the detected and actual reaction events can only be matched if the changes in the pairwise bond orders are similar, which is adopted in the work.

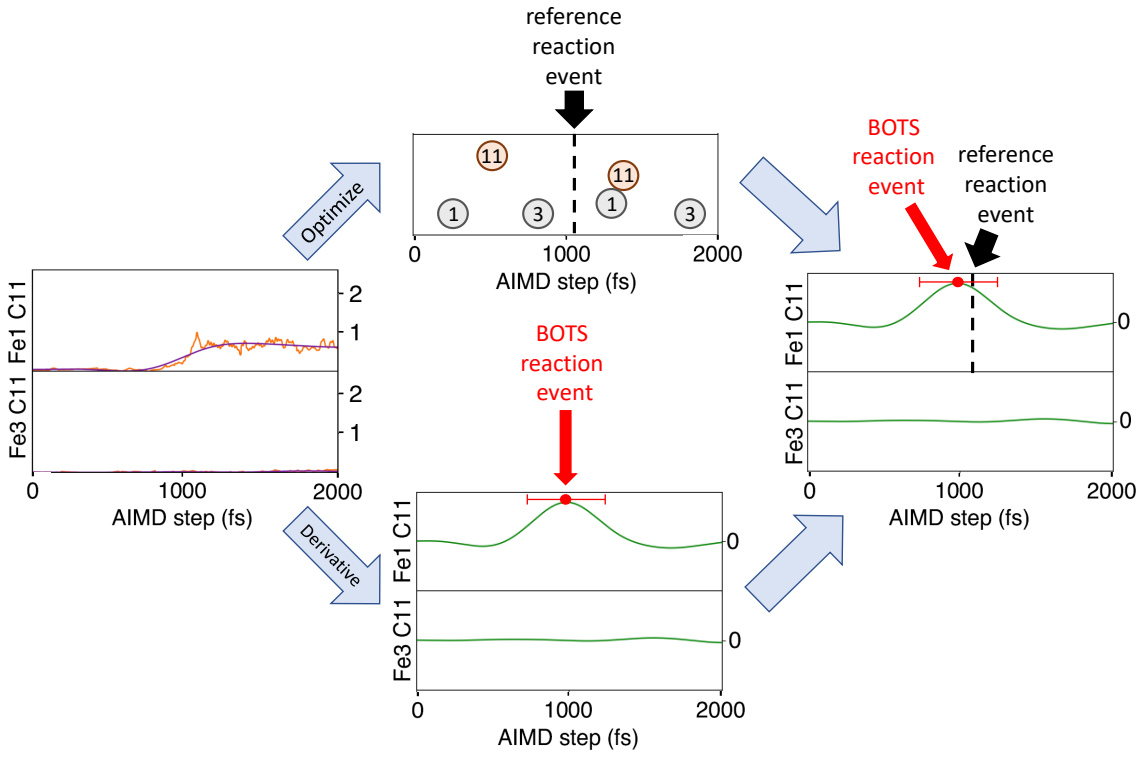

Figure 5: Left: Raw and smoothed bond order time series. Top: Optimized positions of atoms. Reference reaction event near 1000 fs experienced a large change in its optimized BO matrix from atom pair (1, 11). Bottom: BOTS method predicting a reaction event by identifying extrema in the first time derivative of the BO time series beyond a threshold. Right: The BOTS-predicted reaction event in (1, 11) is compared to the reference reaction event for that atom pair.

In the bond-wise ROC, illustrated in Figure 5, predicted reaction events in $\mathcal{E}_{1;ab}$ can only be matched to reference events that involve significant changes in the BO of atom pair

ab. This procedure involves defining the pairwise BO difference between clusters of energy-minimized structures. We first define an averaged BO matrix over the energy-minimized structures within the cluster as:

$$\widehat{M}_{ab}[i] \equiv \frac{1}{|C_{K[i]}|} \sum_{j \in C_{K[i]}} \widetilde{M}_{ab}[j] \tag{16}$$

where K is a cluster index, i, j are frame indices and a, b are atom indices. This enables the definition of absolute pairwise BO difference between clusters as:

$$\Delta_{ab}[i] \equiv \mathrm{abs}(\widehat{M}_{ab}[i+1] - \widehat{M}_{ab}[i]) \tag{17}$$

The reference reaction events involving atom pair (a, b), given by $\mathcal{E}_{\mathrm{ref},ab}$, is defined as:

$$\mathcal{E}_{\mathrm{ref},ab} = \{i \ | \Delta_{ab}[i] \geq 0.5 \max(\boldsymbol{\Delta}[i])\} \cap \mathcal{E}_{\mathrm{ref}} \tag{18}$$

where $\boldsymbol{\Delta}[i]$ is the BO difference matrix between clusters $K[i], K[i+1]$ and the maximum is taken over all pairs of atoms. Thus, each individual event in $\mathcal{E}_{\mathrm{ref}}$ may be included in one or more bond-wise sets $\mathcal{E}_{\mathrm{ref},ab}$.

The trajectory frames being optimized within a time window w of $\mathcal{E}_{1;ab}$ is denoted using $\mathcal{X}_{ab}^{(w)}$ and defined in a similar manner to Eqs.13-14 with $\mathcal{E}_{1;ab}$ replacing $\mathcal{E}_1$. The true positive rate with the added bond-wise criterion is then defined by taking the union over all successfully found reaction events in the numerator:

$$\mathrm{TPR}'^{(w)} = \frac{\left| \bigcup_{b>a=1}^{N_{\mathrm{atoms}}} \left(\mathcal{X}_{ab}^{(w)} \bigcap \mathcal{E}_{\mathrm{ref};ab} \right) \right|}{|\mathcal{E}_{\mathrm{ref}}|} \tag{19}$$

The corresponding false positive rate represents the ratio of all energy-minimized frames not corresponding to reaction events in the numerator, and the same denominator as in Equation 15. Due to the extra condition imposed by the bond-wise criterion, the numerator

may slightly exceed the denominator when $\left|\mathcal{X}^{(w)}\right|$ approaches the trajectory length; the FPR is set equal to 1.0 when this occurs. Additionally, the predicted reaction events from a random number generator no longer result in a ROC of 0.5 due to the additional conditions imposed on matching a reference reaction event to a predicted one.

$$\mathrm{FPR}'^{(w)} = \min\left(\frac{\left|\mathcal{X}^{(w)} - \bigcup_{b>a=1}^{N_{\mathrm{atoms}}}\left(\mathcal{X}_{ab}^{(w)} \bigcap \mathcal{E}_{\mathrm{ref};ab}\right)\right|}{|\mathcal{T} - \mathcal{E}_{\mathrm{ref}}|}, 1\right) \tag{20}$$

Similar to before, the bond-wise ROC $\phi'(\sigma, \mu)$ is calculated as the area under the parametric curve $(\mathrm{FPR}'^{(w)}, \mathrm{TPR}'^{(w)})$. We will drop the primes in the next section, as our results will use the bond-wise criterion exclusively.

3 Results and Discussion

In this section, we characterize the performance and parameter sensitivity of our reaction detection models. The primary means of measuring performance is the ROC $\phi(\sigma, \mu)$ discussed above, and the parameter sensitivity is characterized by observing how the ROC varies with respect to its two parameters: the cutoff frequency in the low pass filter σ (given in cm^{-1}, and the threshold on the time derivative μ given in units of multiples of σ. Because the parameter space is two-dimensional, the global optimum and parameter sensitivity can be obtained by plotting $\phi(\sigma, \mu)$ as a heat map.

3.1 Heptanylium cation

The reaction events observed in the AIMD trajectory for heptanylium cation ($C_7H_{15}{}^+$) mostly involve hydride shifts where H^- is transferred from a non-terminal CH_2 group to the neighboring trivalent carbon with a formal positive charge. The energy-minimized local minima, shown in the top row of Figure 7, include carbocation species with a formally pos-

itive trivalent carbon (clusters 3-6) as well as carbonium species with a pentavalent carbon (clusters 1-2). The heat map for heptanylium in Figure 6 shows that the ROC objective function $\phi(\sigma, \mu)$ has values above 0.95 in a broad region of parameter space, indicating a high degree of accuracy in detecting reaction events that is not highly sensitive to parameter choice. The objective function value indicates that most or all of the predicted events with only small time differences from the reference events. The global optimum is given as $\phi(\sigma = 140\ \text{cm}^{-1}, \mu = 1.0) = 0.97$.

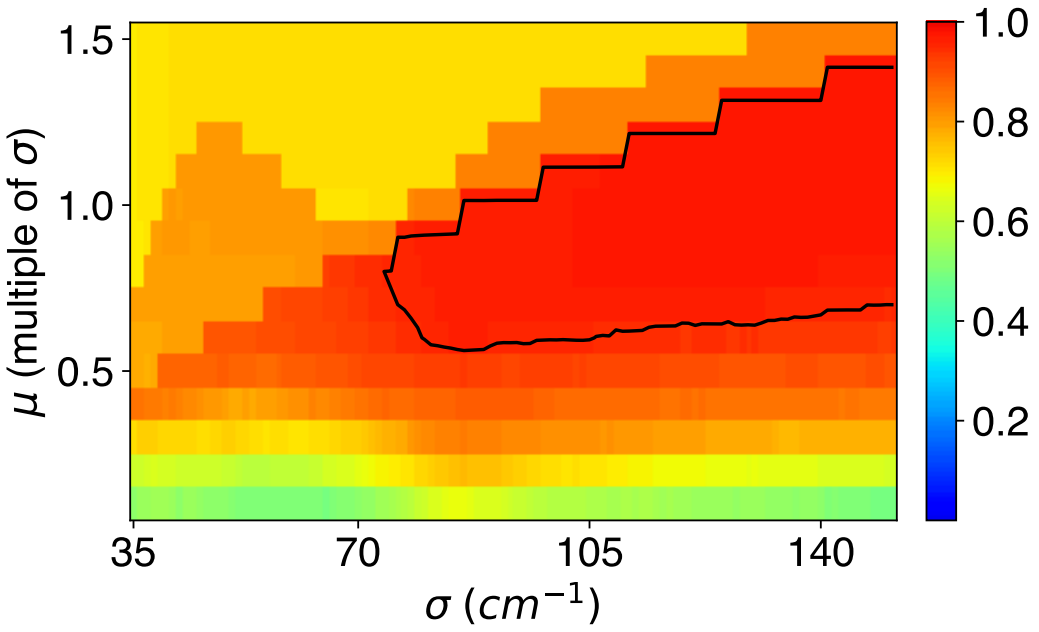

Figure 6: Heat map of bond-wise objective function scores for $C_7H_{15}^+$ using different combinations of σ and μ. Black contour indicates scores above 0.95.

Figure 7 examines the level of agreement between predicted and reference reaction events using the optimal parameter combination of $\sigma = 140cm^{-1}$ and $\mu = 1.0\sigma$ identified from the heat map. There are 17 predicted and 13 reference reaction events respectively, and the maximum time difference between any reference event and the nearest predicted event that satifies the bond-wise criterion was 42 frames. The set of energy-minimized trajectory frames using a window size of 42 ($\mathcal{X}^{(42)}$) covers 7.9% of the whole trajectory, which is another indicator of the accuracy of the reaction detection model. Figure 7 also shows the starting

and ending cluster numbers for each reference reaction event. The colors of arrows indicate the time difference between the reference and predicted reaction events. From this data, we observed that reference reaction events have a tendency to occur in multiplets due to re-crossing of dividing surfaces. Some reference reaction events occur in closely spaced opposite pairs, such as cluster number 6 which is visited once from cluster number 5.

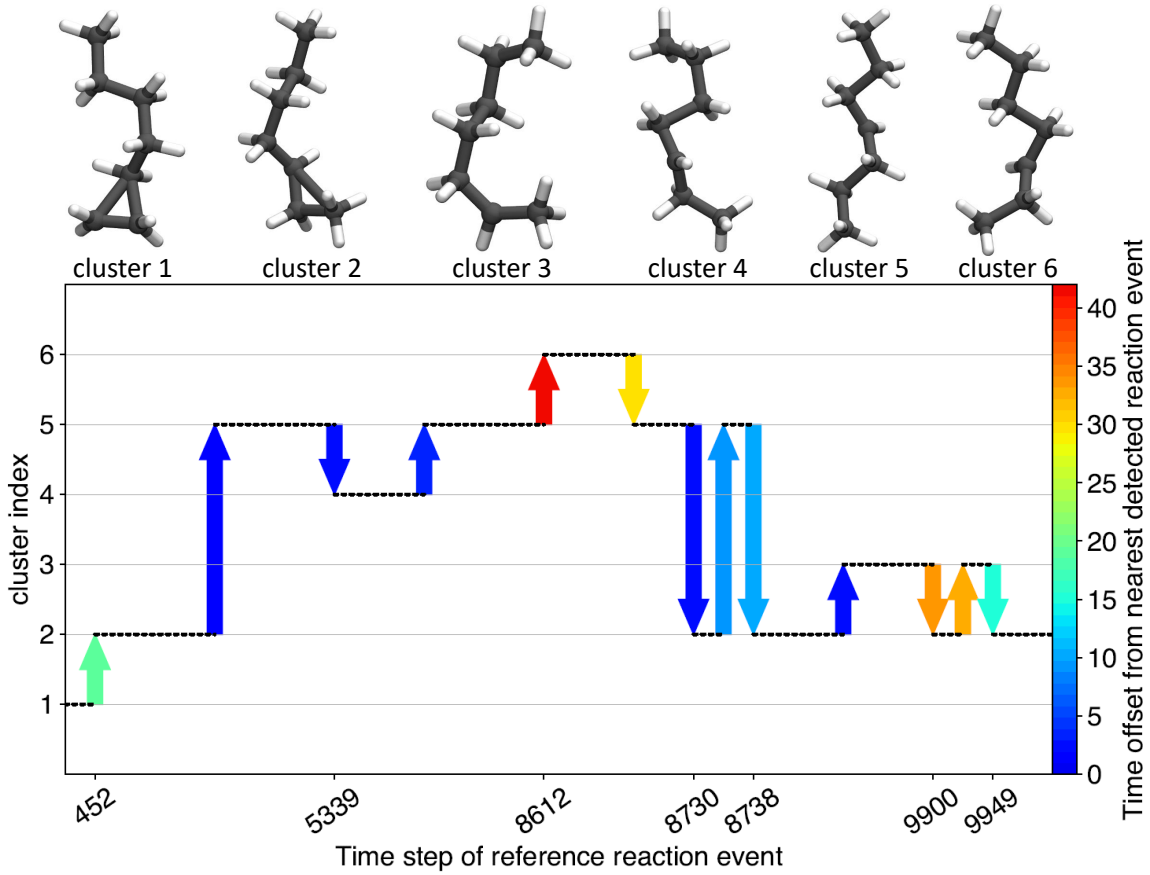

Figure 7: Summary of reaction events and detection model in heptanylium cation ($C_7H_{15}^+$) simulation trajectory. Top: Chemically distinct clusters found after minimizing all trajectory frames followed by clustering. Bottom: Comparison of reaction events for reference and BOTS predictions. Horizontal coordinates of arrows indicate the time step (x axis not to scale), vertical coordinates indicate starting and ending cluster, and color indicates temporal proximity to nearest BOTS prediction. Parameter combination of $\sigma = 140cm^{-1}$ and $\mu = 1.0\sigma$ was chosen from the optimal parameter range shown in Figure 6.

3.2 Iron carbonyl cluster

The AIMD trajectory for the iron carbonyl cluster begins with an optimized structure reported by Schaefer and coworkers,[49] denoted as "9a" in their publication. This system is characterized by nearly constant, almost fluid migration of carbonyl ligands throughout the duration of the simulations, whereas the Fe atoms move more slowly due to their increased mass. The carbonyls migrate by breaking and forming coordinations with individual irons and breaking and forming bridging relationships across multiple irons.

The reactivity in this system is more difficult to characterize compared to the heptanylium system for several reasons. One reason is that the number of chemically distinct clusters and reaction events was simply higher in this trajectory. Perhaps more importantly, the chemical bonding in this system is less discrete compared to the previous case, because the Fe-C and Fe-Fe bond orders of the energy minimized structures are more broadly distributed between 0 and 1. This is also evident in the dendrogram of Figure S2, which shows that the number of clusters and reference reaction events has a significant dependence on the clustering threshold. Thus, this system approaches the limits of our basic assumptions that the potential energy surface consists of discrete and well-separated chemical species.

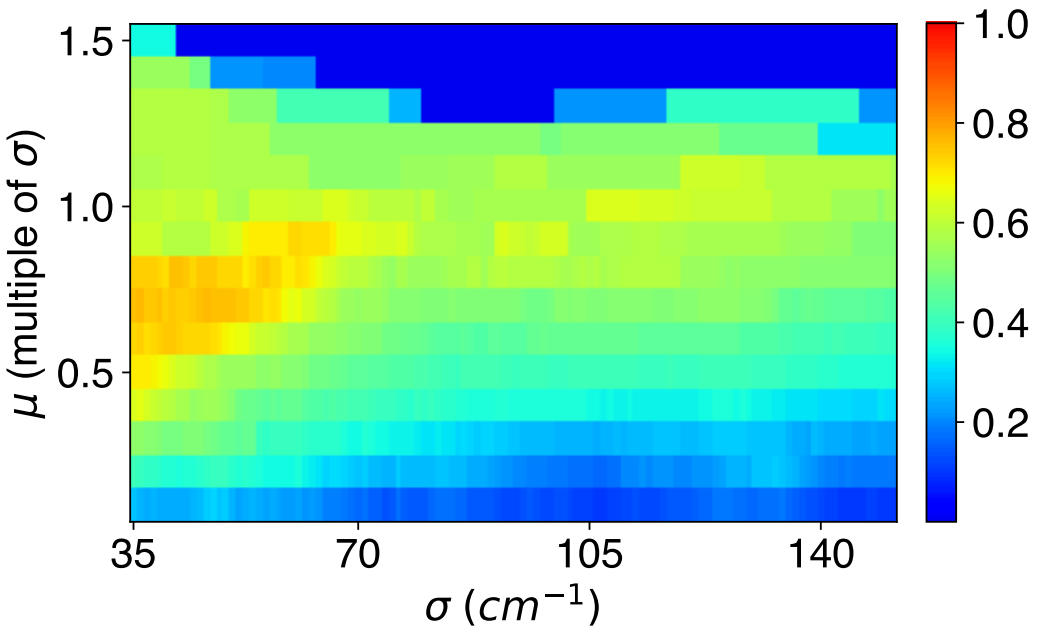

Figure 8: Heat map of ROC objective scores scores for $Fe_3(CO)_9$ as a function of σ and μ.

The heat map for the $Fe_3(CO)_9$ system is shown in Figure 8. Compared to the $C_7H_{15}{}^+$system, the objective function scores are generally lower and there is no parameter combination that gives a score above 0.9, but there still exists a region of parameter space that gives the optimal result as indicated by the orange area. These ideal parameter combinations occur lower σ values and lower μ values than in Figure 6, which we think are due to the slower overall dynamics of the system, owing to the increased mass of Fe and perhaps the relatively flat potential energy surface along reaction coordinates. The difference in optimal parameters between the heptanylium and the iron carbonyl simulation trajectories indicates that this method is not completely system independent. However, it does appear possible to choose parameter sets based on the elemental composition of the system without needing to determine parameters for each individual AIMD trajectory.

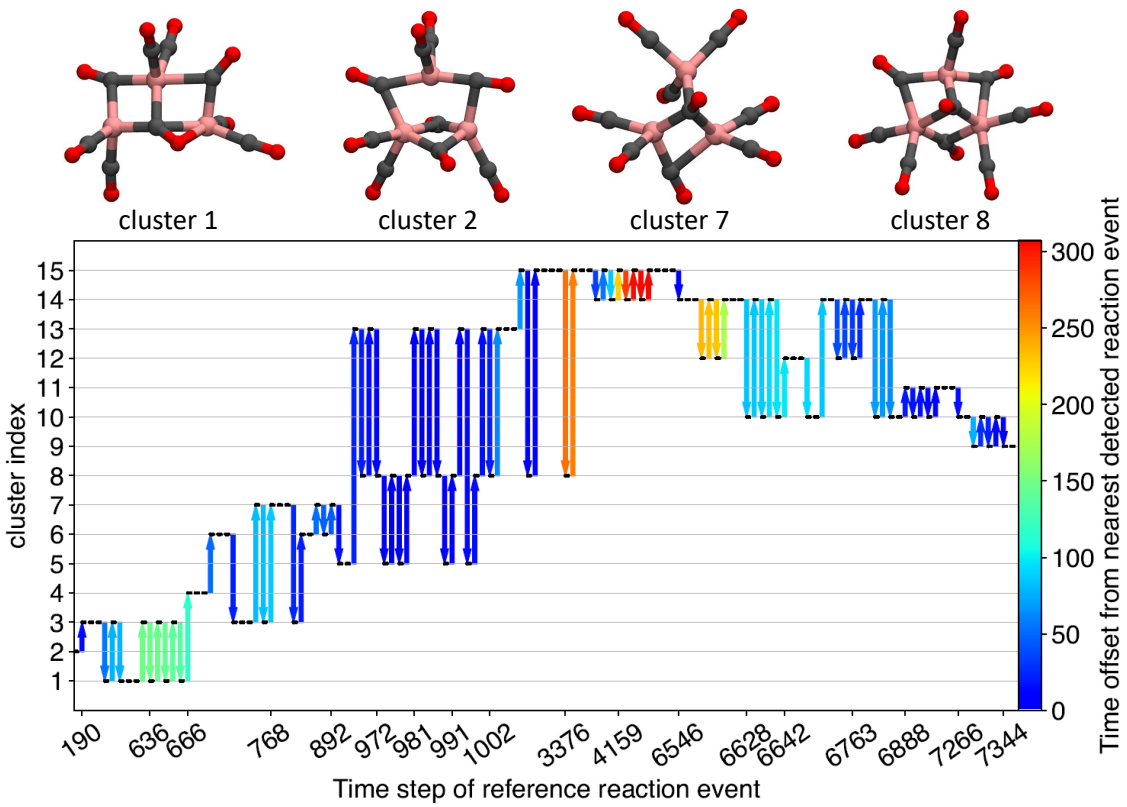

Figure 9: Summary of reaction events and detection model in $Fe_3(CO)_9$ simulation trajectory. Top: Selection of chemically distinct clusters found after minimizing all trajectory frames followed by clustering. All clusters in Supporting Figure S8. Bottom: Comparison of reaction events for reference and BOTS predictions. Horizontal coordinates of arrows indicate the time step (x axis not to scale), vertical coordinates indicate starting and ending cluster, and color indicates temporal proximity to nearest BOTS prediction. Parameter combination of $\sigma = 40cm^{-1}$ and $\mu = 0.7\sigma$ was chosen from the optimal parameter range shown in Figure 8. Molecule color scheme: iron: pink, carbon: gray, and oxygen: red.

A representative parameter set obtained from the optimal range in Figure 8 for $Fe_3(CO)_9$ is given by $\sigma = 40cm^{-1}$ and $\mu = 0.7\sigma$. Using those parameters, Figure 9 shows the temporal proximity of reference reaction events to the nearest BOTS predictions that satisfy the bond-wise criterion. The data shows that reference reaction events have a strong tendency to be grouped together as the dividing surface is crossed multiple times within a short simulation time. There is also a large variation in the "difficulty" of detecting certain reaction events vs. other ones, as indicated by the temporal distance between the reference reaction event and

the closest detected event. If a window size of 150 frames is used, 87% of reference reaction events can be found, which would require energy-minimizing 50% of the trajectory frames. To find the remaining 13% of reference reaction events in this trajectory, the time window needs to be 310 frames, which covers 68% of the trajectory. Closer inspection of the most difficult reference reaction events reveals that they occur in closely spaced opposite pairs, where the cluster index jumps to a new value for ∼10 frames then back again. Thus, we think that for challenging systems such as these, it may not be necessary to find 100% of the reaction events in order to get a comprehensive picture of the reactivity of the system. In applications where computational cost is a critical concern, the reference reaction events may be found more quickly (if not as thoroughly) using other methods such as skipping frames when extending the window, that may be more relevant as post-processing approaches than objective functions. In this context, the objective function score should not be seen as a literal measure of computational cost savings, but rather as a measure of the accuracy of reaction event detection.

4 Conclusion

This paper describes how the time series analysis of bond orders is able to produce accurate predictions of the spatial and temporal locations of reaction events in reactive *ab initio* molecular dynamics trajectories. Reaction events in simple systems like hydrocarbons can be predicted with great accuracy; more complex and fluxional systems like iron carbonyl clusters contain reaction events that may still be identified, though not as easily. The accuracy of reaction event prediction can translate into more efficient computations, as it reduces the portions of the simulation trajectory that need to be examined in greater detail using methods such as geometry optimization. Our reaction detection method contains two adjustable parameters that are not fully system independent, but the optimized parameters of a system are expected to be broadly useful for simulations of chemically similar systems.

A natural extension of this research would be to identify reaction events in multi-molecular simulations in a more rigorous manner. The challenges to be addressed include how to identify the subset of atoms in the overall system that are involved in a given reaction event, which could also be informed by analysis of the bond order matrix. Because the bond order matrix contains rich information about the chemical structure of the system, it might also be a useful collective variable for future metadynamics or other enhanced-sampling simulations to rapidly explore the chemical space. We anticipate that the bond order matrix will play an increasingly important role in reaction discovery as these methods continue to be developed and applied to chemical problems.

5 Acknowledgements

We would like to acknowledge the ACS-PRF award 58158-DNI6 and the NSF ChemEnergy REU site, award CHE-1560479. C.S. is grateful for an E. K. Potter Stanford Graduate Fellowship and support through NSF ACI-1450179.

References

(1) Soriano, E.; Marco-Contelles, J. Mechanistic Insights on the Cycloisomerization of Polyunsaturated Precursors Catalyzed by Platinum and Gold Complexes. *Acc. Chem. Res.* **2009**, *42*, 1026–1036.

(2) Sletten, E. M.; Bertozzi, C. R. From Mechanism to Mouse: A Tale of Two Bioorthogonal Reactions. *Acc. Chem. Res.* **2011**, *44*, 666–676.

(3) Szymkuc, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chem. Int. Edit.* **2016**, *55*, 5904–5937.

(4) Peng, C.; Ayala, P.; Schlegel, H.; Frisch, M. Using redundant internal coordinates to optimize equilibrium geometries and transition states. *J. Comput. Chem.* **1996**, *17*, 49–56.

(5) Henkelman, G.; Uberuaga, B.; Jonsson, H. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *J. Chem. Phys.* **2000**, *113*, 9901–9904.

(6) Mebel, A.; Diau, E.; Lin, M.; Morokuma, K. Ab initio and RRKM calculations for multichannel rate constants of the C2H3+O-2 reaction. *J. Am. Chem. Soc.* **1996**, *118*, 9759–9771.

(7) Skulason, E.; Tripkovic, V.; Bjorketun, M. E.; Gudmundsdottir, S.; Karlberg, G.; Rossmeisl, J.; Bligaard, T.; Jonsson, H.; Norskov, J. K. Modeling the Electrochemical Hydrogen Oxidation and Evolution Reactions on the Basis of Density Functional Theory Calculations. *J. Phys. Chem. C* **2010**, *114*, 18182–18197.

(8) Habershon, S. Automated Prediction of Catalytic Mechanism and Rate Law Using Graph-Based Reaction Path Sampling. *J. Chem. Theory Comput.* **2016**, *12*, 1786–1798.

(9) Ismail, I.; Stuttaford-Fowler, H. B.; Ochan Ashok, C.; Robertson, C.; Habershon, S. Automatic Proposal of Multistep Reaction Mechanisms using a Graph-Driven Search. *J. Phys. Chem. A* **2019**, *123*, 3407–3417.

(10) Zimmerman, P. M. Automated discovery of chemically reasonable elementary reaction steps. *J. Comput. Chem* **2013**, *34*, 1385–1392.

(11) Dewyer, A. L.; Arguelles, A. J.; Zimmerman, P. M. Methods for exploring reaction space in molecular systems. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2018**, *8*, e1354.

(12) Kim, Y.; Kim, J. W.; Kim, Z.; Kim, W. Y. Efficient prediction of reaction paths through molecular graph and reaction network analysis. *Chem. Sci.* **2018**, *9*, 825–835.

(13) Rappoport, D.; Galvin, C. J.; Zubarev, D. Y.; Aspuru-Guzik, A. Complex chemical reaction networks from heuristics-aided quantum chemistry. *J. Chem. Theory Comput.* **2014**, *10*, 897–907.

(14) Gao, C. W.; Allen, J. W.; Green, W. H.; West, R. H. Reaction Mechanism Generator: Automatic construction of chemical kinetic mechanisms. *Comput. Phys. Commun.* **2016**, *203*, 212–225.

(15) Unsleber, J. P.; Reiher, M. The Exploration of Chemical Reaction Networks. arXiv:1906.10223 [physics.chem-ph], 2019.

(16) Habershon, S. Sampling reactive pathways with random walks in chemical space: Applications to molecular dissociation and catalysis. *J. Chem. Phys.* **2015**, *143*, 094106.

(17) Pendleton, I. M.; Perez-Temprano, M. H.; Sanford, M. S.; Zimmerman, P. M. Experimental and Computational Assessment of Reactivity and Mechanism in C(sp(3))-N Bond-Forming Reductive Elimination from Palladium(IV). *J. Am. Chem. Soc.* **2016**, *138*, 6049–6060.

(18) Rappoport, D.; Aspuru-Guzik, A. Predicting Feasible Organic Reaction Pathways Using Heuristically Aided Quantum Chemistry. *J. Chem. Theory Comput.* **2019**, *15*, 4099–4112.

(19) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

(20) Onuchic, J.; LutheySchulten, Z.; Wolynes, P. Theory of protein folding: The energy landscape perspective. *Annu. Rev. Phys. Chem.* **1997**, *48*, 545–600.

(21) Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. Molecular Simulation of ab Initio Protein Folding for a Millisecond Folder NTL9(1-39). *J. Am. Chem. Soc.* **2010**, *132*, 1526+.

(22) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How Fast-Folding Proteins Fold. *Science* **2011**, *334*, 517–520.

(23) Lapidus, L. J.; Acharya, S.; Schwantes, C. R.; Wu, L.; Shukla, D.; King, M.; Decamp, S. J.; Pande, V. S. Complex pathways in folding of protein G explored by simulation and experiment. *Biophys. J.* **2014**, *107*, 947–955.

(24) Ufimtsev, I. S.; Martínez, T. J. Quantum chemistry on graphical processing units. 1. strategies for two-electron integral evaluation. *J. Chem. Theory Comput.* **2008**, *4*, 222–231.

(25) Ufimtsev, I. S.; Martínez, T. J. Quantum chemistry on graphical processing units. 2. direct self-consistent-field implementation. *J. Chem. Theory Comput.* **2009**, *5*, 1004–1015.

(26) Ufimtsev, I. S.; Martínez, T. J. Quantum chemistry on graphical processing units. 3. Analytical energy gradients, geometry optimization, and first principles molecular dynamics. *J. Chem. Theory Comput.* **2009**, *5*, 2619–2628.

(27) Stone, J. E.; Hardy, D. J.; Ufimtsev, I. S.; Schulten, K. GPU-accelerated molecular modeling coming of age. *J. Mol. Graph. Model.* **2010**, *29*, 116–125.

(28) Titov, A. V.; Ufimtsev, I. S.; Luehr, N.; Martinez, T. J. Generating Efficient Quantum Chemistry Codes for Novel Architectures. *J. Chem. Theory Comput.* **2013**, *9*, 213–221.

(29) Vogt, L.; Olivares-Amaya, R.; Kermes, S.; Shao, Y.; Amador-Bedolla, C.; Aspuru-Guzik, A. Accelerating resolution-of-the-identity second-order Moller-Plesset quantum

chemistry calculations with graphical processing units. *J. Phys. Chem. A* **2008**, *112*, 2049–2057.

(30) Genovese, L.; Ospici, M.; Deutsch, T.; Mehaut, J.-F.; Neelov, A.; Goedecker, S. Density functional theory calculation on many-cores hybrid central processing unit-graphic processing unit architectures. *J. Chem. Phys.* **2009**, *131*.

(31) Yasuda, K. Accelerating density functional calculations with graphics processing unit. *J. Chem. Theory Comput.* **2008**, *4*, 1230–1236.

(32) DePrince, A. E., III; Hammond, J. R. Coupled Cluster Theory on Graphics Processing Units I. The Coupled Cluster Doubles Method. *J. Chem. Theory Comput.* **2011**, *7*, 1287–1295.

(33) Wu, X.; Koslowski, A.; Thiel, W. Semiempirical Quantum Chemical Calculations Accelerated on a Hybrid Multicore CPU-GPU Computing Platform. *J. Chem. Theory Comput.* **2012**, *8*, 2272–2281.

(34) Hacene, M.; Anciaux-Sedrakian, A.; Rozanska, X.; Klahr, D.; Guignon, T.; Fleurat-Lessard, P. Accelerating VASP electronic structure calculations using graphic processing units. *J. Comput. Chem.* **2012**, *33*, 2581–2589.

(35) Song, C.; Wang, L.-P.; Sachse, T.; Preiss, J.; Presselt, M.; Martinez, T. J. Efficient implementation of effective core potential integrals and gradients on graphical processing units. *J. Chem. Phys.* **2015**, *143*.

(36) Wang, L.-P.; Titov, A.; McGibbon, R.; Liu, F.; Pande, V. S.; Martínez, T. J. Discovering chemistry with an ab initio nanoreactor. *Nat. Chem.* **2014**, *6*, 1044–1046.

(37) Goldman, N.; Reed, E. J.; Fried, L. E.; William Kuo, I. F.; Maiti, A. Synthesis of glycine-containing complexes in impacts of comets on early Earth. *Nat. Chem.* **2010**, *2*, 949–954.

(38) Saitta, A. M.; Saija, F. Miller experiments in atomistic computer simulations. *P. Natl. Acad. Sci. USA* **2014**, *111*, 13768–13773.

(39) Pérez-Villa, A.; Saitta, A. M.; Georgelin, T.; Lambert, J. F.; Guyot, F.; Maurel, M. C.; Pietrucci, F. Synthesis of RNA Nucleotides in Plausible Prebiotic Conditions from ab Initio Computer Simulations. *J. Phys. Chem. Lett.* **2018**, *9*, 4981–4987.

(40) Martínez-Núñez, E. An automated method to find transition states using chemical dynamics simulations. *J. Comput. Chem.* **2015**, *36*, 222–234.

(41) Rodríguez, A.; Rodríguez-Fernández, R.; A. Vázquez, S.; L. Barnes, G.; J. P. Stewart, J.; Martínez-Núñez, E. tsscds2018: A code for automated discovery of chemical reaction mechanisms and solving the kinetics. *J. Comput. Chem.* **2018**, *39*, 1922–1930.

(42) Pietrucci, F.; Andreoni, W. Graph theory meets ab initio molecular dynamics: Atomic structures and transformations at the nanoscale. *Phys. Rev. Lett.* **2011**, *107*.

(43) Fu, C. D.; Pfaendtner, J. Lifting the Curse of Dimensionality on Enhanced Sampling of Reaction Networks with Parallel Bias Metadynamics. *J. Chem. Theory Comput.* **2018**, *14*, 2516–2525.

(44) Das, T.; Ghule, S.; Vanka, K. Insights Into the Origin of Life: Did It Begin from HCN and H2O? *ACS Cent. Sci.* **2019**, *5*, 1532–1540.

(45) Meisner, J.; Zhu, X.; Martínez, T. J. Computational Discovery of the Origins of Life. *ACS Cent. Sci.* **2019**, *5*, 1493–1495.

(46) Wang, L. P.; McGibbon, R. T.; Pande, V. S.; Martinez, T. J. Automated Discovery and Refinement of Reactive Molecular Dynamics Pathways. *J. Chem. Theory Comput.* **2016**, *12*, 638–649.

(47) Döntgen, M.; Przybylski-Freund, M.-D.; Kröger, L. C.; Kopp, W. A.; Ismail, A. E.; Leonhard, K. Automated Discovery of Reaction Pathways, Rate Constants, and Transi-

tion States Using Reactive Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2015**, *11*, 2517–2524, PMID: 26575551.

(48) Mayer, I. Bond order and valence indices: A personal account. *J. Comput. Chem.* **2007**, *28*, 204–221.

(49) Wang, H.; Xie, Y.; King, R. B.; Schaefer, H. F. Remarkable Aspects of Unsaturation in Trinuclear Metal Carbonyl Clusters: The Triiron Species Fe3(CO)n (n = 12, 11, 10, 9). *J. Am. Chem. Soc.* **2006**, *128*, 11376–11384, PMID: 16939260.

(50) Sorensen, M.; Voter, A. Temperature-accelerated dynamics for simulation of infrequent events. *J. Chem. Phys.* **2000**, *112*, 9599–9606.

(51) Xie, L.; Zhao, Q.; Jensen, K. F.; Kulik, H. J. Direct Observation of Early-Stage Quantum Dot Growth Mechanisms with High-Temperature Ab Initio Molecular Dynamics. *J. Phys. Chem. C* **2016**, *120*, 2472–2483.

(52) Wang, L. P.; Song, C. Geometry optimization made simple with translation and rotation coordinates. *J. Chem. Phys.* **2016**, *144*.

(53) Harrigan, M. P.; Sultan, M. M.; Hernández, C. X.; Husic, B. E.; Eastman, P.; Schwantes, C. R.; Beauchamp, K. A.; McGibbon, R. T.; Pande, V. S. MSMBuilder: Statistical Models for Biomolecular Dynamics. *Biophys. J.* **2017**, *112*, 10 – 15.

(54) Husic, B. E.; Pande, V. S. Markov State Models: From an Art to a Science. *J. Am. Chem. Soc.* **2018**, *140*, 2386–2396, PMID: 29323881.

(55) Jones, E.; Oliphant, T.; Peterson, P. SciPy: Open source scientific tools for Python. `http://www.scipy.org/`.

(56) Bar-Joseph, Z.; Gifford, D. K.; Jaakkola, T. S. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics* **2001**, *17*, S22–S29.

(57) Hanley, J. A.; McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36.

(58) Fawcett, T. An introduction to ROC analysis. *Pattern Recogn. Lett.* **2006**, *27*, 861–874.